Animaltracker Data Validation: New Mexico Data

Joe Champion, Thea Sukianto

May 22, 2020

This document analyzes the results of the `animaltracker` package's data cleaning procedures by comparing data flagged by the app to data flagged by manual processing via spreadsheet.

The cleaning process uses flag-based rules for discarding cases (rows) of data.

- If the `Rate` > 84, mark the case with a `RateFlag`.
- If the `Course` ≥ 100, mark the case with a `CourseFlag`.
- If the `Distance` ≥ 840, mark the case with a `DistanceFlag`.
- Discard any case with a `DistanceFlag`, or 2+ flags (or both).

Preliminaries

Configure and load needed packages (use `install.packages("packagename")` to install any missing libraries).

```r
library(dplyr)
library(ggplot2)
library(tidyr)
```

Prepare Data

```r
clean_anitracker <- read.csv("df_candidate.csv", stringsAsFactors = FALSE) %>%
  #################
  ### !!! HOT FIX FOR ERROR IN GEODIST
  ## IMPLEMENT IN APP, THEN DELETE AFTER RE-CLEANING
  mutate(
    DistGeo = ifelse(DistGeo < 10^6, DistGeo, 0), ### !!! hot fix for GeoDist error
    Rate = ifelse(TimeDiffMins != 0, DistGeo/TimeDiffMins, 0),
    RateFlag = 1*(Rate > 84),
    DistanceFlag = 1*(DistGeo >= 840)
  )
  #################
clean_manual <- read.csv("df_correct.csv", stringsAsFactors = FALSE)
```

First, we join the cleaned data from the animaltracker app (167901 rows, 36 columns) with the cleaned data from manual processing (167901 rows, 31 columns).

Rows are matched by the combination of `Cow`, `Index` (uniquely identifies almost all rows) and `Altitude` (to break ties in rare duplicates).

```r
clean_anitracker <- clean_anitracker %>%
  arrange(Cow, Index, Altitude) %>%
  mutate(merge_index = 1:n())
```

```
clean_manual <- clean_manual %>%
  arrange(Cow, Index, Altitude) %>%
  mutate(merge_index = 1:n())

join <- dplyr::full_join(clean_anitracker, clean_manual, by="merge_index") %>%
  dplyr::rename(Index = Index.y,
                Cow = Cow.y,
                Altitude = Altitude.y,
                Order = Order.y,
                Keep.y = Keep,
                Speed = Speed.x,
                Course = Course.x,
                DateTime = DateTime.x,
                Dist.x = Distance.x,
                Dist.y = Distance.y,
                DistFlag.x = DistanceFlag,
                DistFlag.y = DistFlag) %>%
  dplyr::mutate(Keep.x = 1*(TotalFlags.x < 2 & !DistFlag.x))
```

The merged data has the 167901 rows.

Analysis

Overall Agreement

First, we compare the results of cleaning the data within `animaltracker` (via the `clean_location_data` function) to results of manual cleaning via spreadsheet.

```
keepxtab <- with(join, table(Keep.x, Keep.y))
```

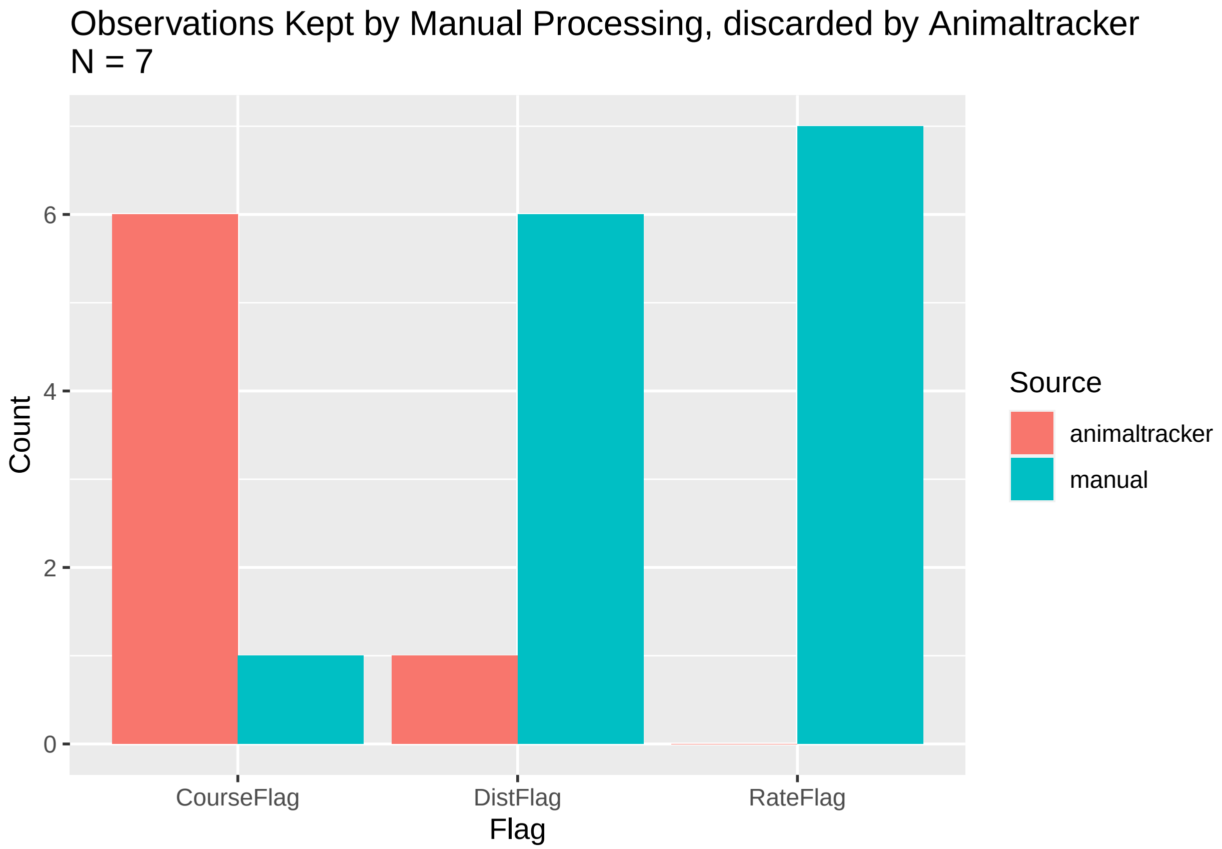
The cleaning methods agree in 99.85% of cases, except for 242 cases (0.14%) kept by `animaltracker` but discarded by manual processing and 7 cases (0%) kept by manual processing but discarded by `animaltracker`.

Analysis of Cases with Different Results

All cases kept by manual processing (n = 7) but discarded by `animaltracker` were marked with a `RateFlag` by manual, but not animaltracker.

```
manual_keep <- join %>%
  dplyr::filter(Keep.x < Keep.y) %>%
dplyr::select(ind = merge_index, Cow, DateTime, Speed, Course, TimeDiffMins, Rate.x, Dist.x, Rate.y, Di

manual_keep %>%
  dplyr::summarise(RateFlag.x = sum(RateFlag.x),
                   CourseFlag.x = sum(CourseFlag.x),
                   DistFlag.x = sum(DistFlag.x),
                   RateFlag.y = sum(RateFlag.y),
                   CourseFlag.y = sum(CourseFlag.y),
                   DistFlag.y = sum(DistFlag.y)) %>%
  tidyr::gather("Flag", "Count") %>%
  dplyr::mutate(Source = ifelse(grepl(".x", Flag), "animaltracker", "manual"),
                Flag = substr(Flag, 1, nchar(Flag)-2)) %>%
  ggplot( aes(Flag, Count, fill = Source)) +
  geom_bar(stat = "identity", position = "dodge") +
```

```r
  ggtitle(paste0("Observations Kept by Manual Processing, discarded by Animaltracker\n","N = ",nrow(manu
```

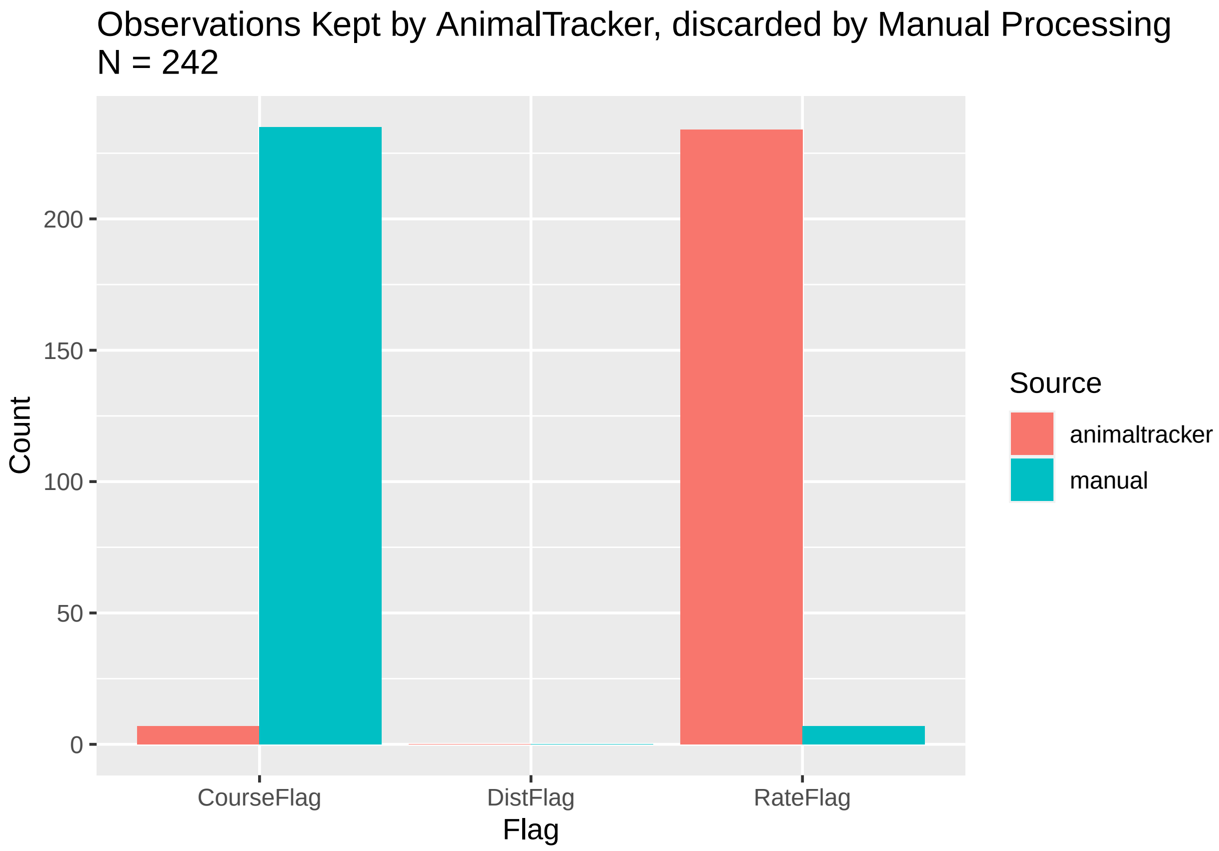


```r
manual_keep %>% head(10)
```

```
##      ind Cow            DateTime Speed Course  TimeDiffMins    Rate.x
## 1  68236 229 2018-05-23 15:45:32     0    239    0.10000000    0.0000
## 2  68272 229 2018-06-12 18:03:18  3168    121    0.00000000    0.0000
## 3  75624 257 2018-05-23 15:31:33     0    184    0.10000000    0.0000
## 4  99860 322 2018-05-23 15:12:19     0    187    0.18333330    0.0000
## 5  99906 322 2018-05-23 16:39:35     0      0    0.78333330    0.0000
## 6  99907 322 2018-05-23 16:39:40     0    303    0.08333333    0.0000
## 7 119295 437 2018-05-23 15:27:10   900     36 -669.40000000 -806.5503
##         Dist.x      Rate.y       Dist.y RateFlag.x CourseFlag.x DistFlag.x
## 1 0.000000e+00           0 0.000000e+00          0            1          0
## 2 5.903604e+05        <NA> 5.903604e+05          0            1          0
## 3 0.000000e+00           0 0.000000e+00          0            1          0
## 4 0.000000e+00           0 0.000000e+00          0            1          0
## 5 0.000000e+00           0 0.000000e+00          0            1          0
## 6 3.019646e+00 36.23575264 3.019646e+00          0            1          0
## 7 0.000000e+00        <NA> 0.000000e+00          0            0          1
##   RateFlag.y CourseFlag.y DistFlag.y
## 1          1            0          1
## 2          1            0          0
## 3          1            0          1
## 4          1            0          1
## 5          1            0          1
## 6          1             0           1
## 7          1             1           1
```

Nearly all cases kept by `animaltracker` but discarded by manual processing (n = 242) had different values of `RateFlag` and `CourseFlag`.

```
anitracker_keep <- join %>%
  dplyr::filter(Keep.x > Keep.y) %>%
  dplyr::select(ind = merge_index, Cow, DateTime, Speed, Course, TimeDiffMins, Rate.x, Dist.x, Rate.y, I

anitracker_keep %>%
  dplyr::summarise(RateFlag.x = sum(RateFlag.x),
                   CourseFlag.x = sum(CourseFlag.x),
                   DistFlag.x = sum(DistFlag.x),
                   RateFlag.y = sum(RateFlag.y),
                   CourseFlag.y = sum(CourseFlag.y),
                   DistFlag.y = sum(DistFlag.y)) %>%
  tidyr::gather("Flag", "Count") %>%
  dplyr::mutate(Source = ifelse(grepl(".x", Flag), "animaltracker", "manual"),
                Flag = substr(Flag, 1, nchar(Flag)-2)) %>%
  ggplot( aes(Flag, Count, fill = Source)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle(paste0("Observations Kept by AnimalTracker, discarded by Manual Processing\n","N = ",nrow(ani
```



```
anitracker_keep %>% head(10)
```

```
##     ind Cow            DateTime Speed Course TimeDiffMins    Rate.x   Dist.x
## 1    57  11 2018-05-23 17:35:39 17892    163     2.116667 101.71443 215.7921
## 2    63  11 2018-05-23 17:48:27 25452    181     2.133333 342.54706 730.1500
## 3    93  11 2018-05-23 18:52:50  5148     96     2.166667 356.48408 770.9845
## 4    99  11 2018-05-23 19:05:50     0    285     2.166667 278.54057 604.4719
## 5   106  11 2018-05-23 19:21:00 52848    147     2.150000 268.46309 578.2565
## 6  1562  11 2018-05-25 20:43:02  8136    315     2.083333  85.65587 178.3132
## 7  1569  11 2018-05-25 20:57:37     0    359     2.100000 108.68097 228.8065
## 8  1575  11 2018-05-25 21:10:07  4212    339     2.100000  88.68540 186.6357
## 9  1579  11 2018-05-25 21:18:29  3960    348     2.100000 112.80285 237.4840
## 10 3636  11 2018-05-28 18:05:02     0    190     2.083333  84.73378 176.8508
##         Rate.y   Dist.y RateFlag.x CourseFlag.x DistFlag.x RateFlag.y
## 1  101.9490098 215.7921          1            0          0          0
## 2  342.2578265 730.1500          1            0          0          0
## 3  355.8389979 770.9845          1            0          0          0
## 4   278.987009 604.4719          1            0          0          0
## 5  268.9564973 578.2565          1            0          0          0
## 6  85.59034731 178.3132          1            0          0          0
## 7  108.9554663 228.8065          1            0          0          0
## 8  88.87416245 186.6357          1            0          0          0
## 9  113.0876033 237.4840          1            0          0          0
## 10 84.88838091 176.8508          1            0          0          0
##    CourseFlag.y DistFlag.y
## 1             1          0
## 2             1          0
## 3             1          0
## 4             1          0
## 5             1          0
## 6             1          0
## 7             1          0
## 8             1          0
## 9             1          0
## 10            1          0
```

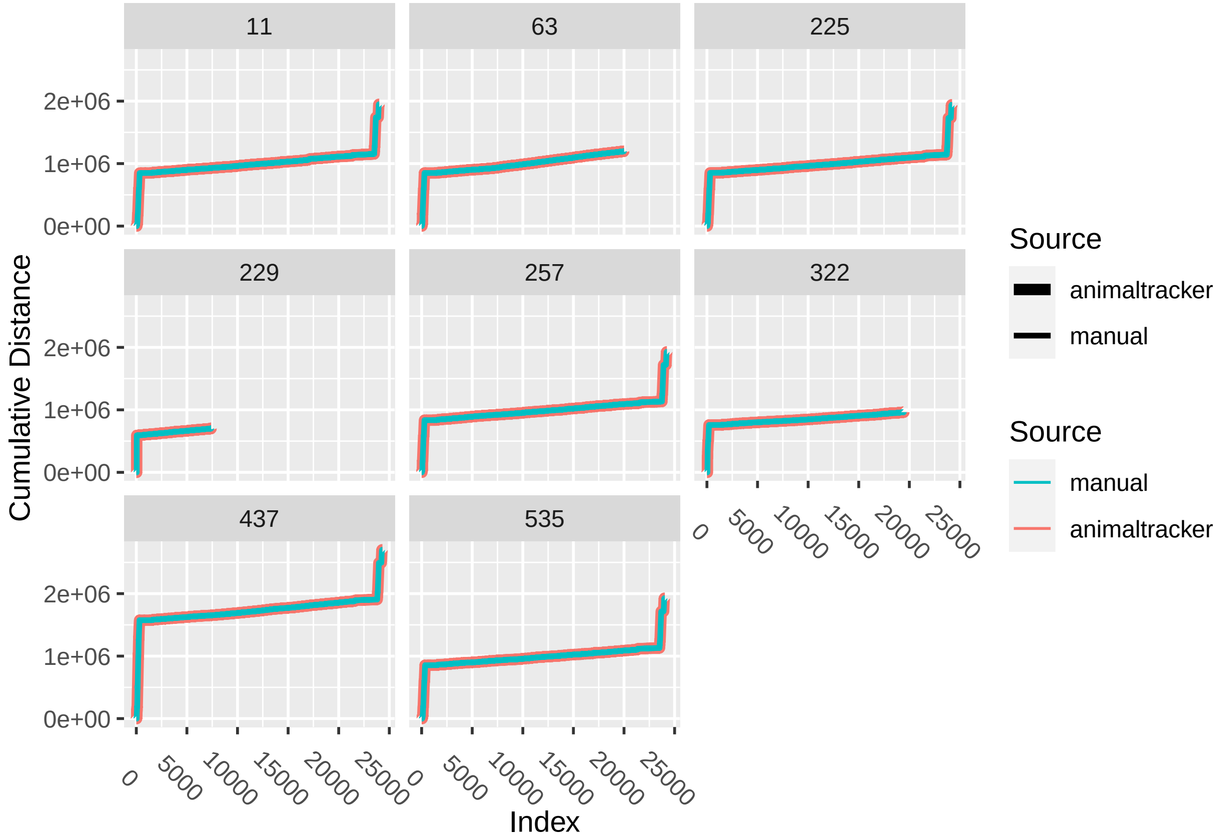
Effects of Cleaning Differences on Outcome Measures

As evidenced by the split time series plots below, there are no substantive differences between the cleaned datasets in cumulative distances, `Rate`, or `Course`.

Cumulative Distance by Cow

```r
cumdist <- join %>%
  dplyr::group_by(Cow) %>%
  dplyr::arrange(Index, .by_group=TRUE) %>%
  dplyr::mutate(Dist.y = dplyr::lag(Dist.y,1),
                Dist.x = ifelse(is.na(Dist.x), 0, Dist.x),
                Dist.y = ifelse(is.na(Dist.y), 0, Dist.y),
                cumDist.x = cumsum(Dist.x),
                cumDist.y = cumsum(Dist.y)) %>%
  dplyr::ungroup()

cumdist_anitracker <- cumdist %>%
  dplyr::select(Index, Cow, cumDist.x, DistFlag.x) %>%
  dplyr::rename(Flag = DistFlag.x,
                cumDist = cumDist.x) %>%
  dplyr::mutate(Source = "animaltracker")
```

```
cumdist_manual <- cumdist %>%
  dplyr::select(Index, Cow, cumDist.y, DistFlag.y) %>%
  dplyr::rename(Flag = DistFlag.y,
                cumDist = cumDist.y) %>%
  dplyr::mutate(Source = "manual")

plot_data <- dplyr::bind_rows(cumdist_anitracker, cumdist_manual)

ggplot(plot_data, aes(x=Index, y=cumDist, group=Source, color=Source)) +
  geom_line(aes(size = Source)) +
  ylab("Cumulative Distance") +
  scale_color_discrete(guide = guide_legend(reverse = TRUE)) +
  scale_size_manual(values=c(2, 1)) +
  facet_wrap(vars(Cow)) +
  theme(axis.text.x = element_text(angle = -45))
```



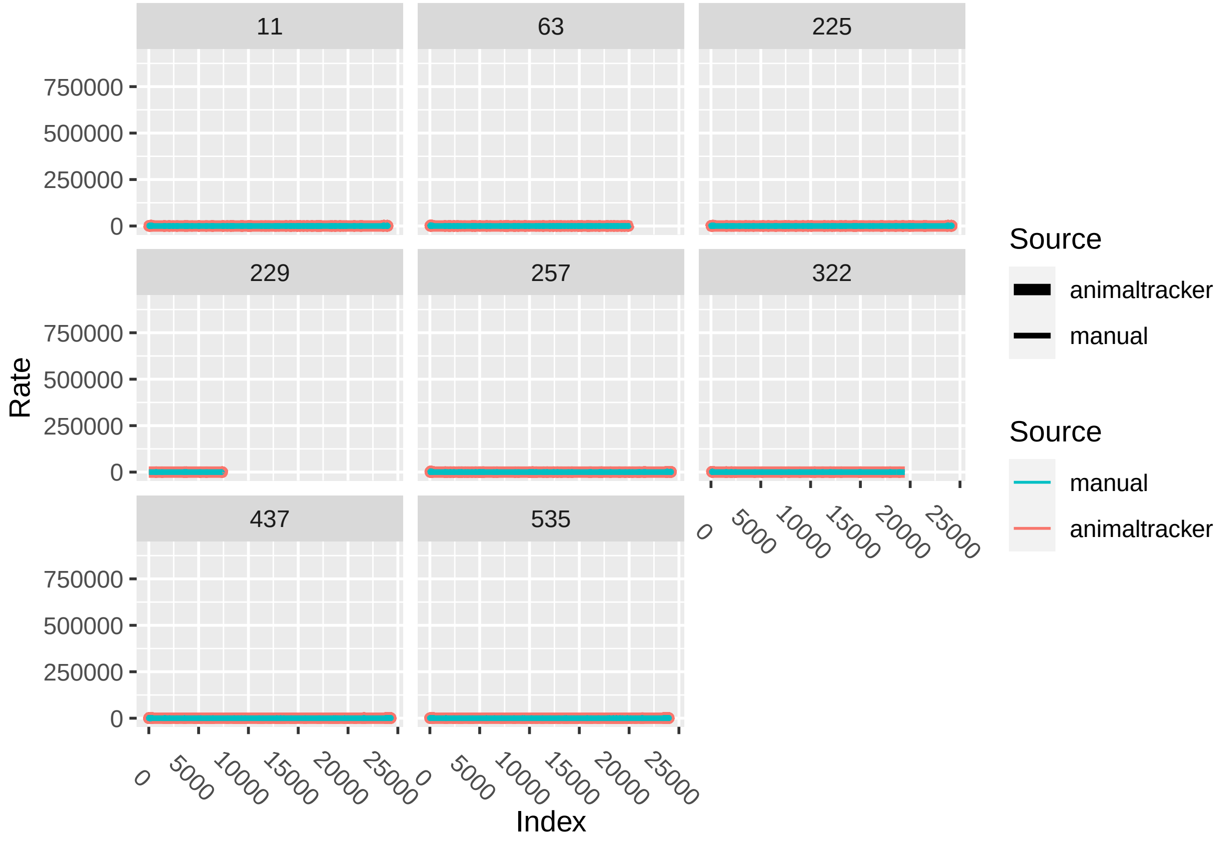
Rate by Cow

```
rate_anitracker <- join %>%
  dplyr::select(Index, Cow, Rate.x, RateFlag.x) %>%
  dplyr::rename(Flag = RateFlag.x,
                Rate = Rate.x) %>%
  dplyr::mutate(Source = "animaltracker")

rate_manual <- join %>%
  dplyr::select(Index, Cow, Rate.y, RateFlag.y) %>%
  dplyr::mutate(Flag = RateFlag.y,
                Rate = as.numeric(Rate.y)) %>%
  dplyr::mutate(Source = "manual")
```

```
## Warning: NAs introduced by coercion
```

```
plot_data <- dplyr::bind_rows(rate_anitracker, rate_manual)

ggplot(plot_data, aes(x=Index, y=Rate, group=Source, color=Source)) +
  geom_line(aes(size = Source)) +
  ylab("Rate") +
  scale_color_discrete(guide = guide_legend(reverse = TRUE)) +
  scale_size_manual(values=c(2, 1)) +
  facet_wrap(vars(Cow)) +
  theme(axis.text.x = element_text(angle = -45))
```

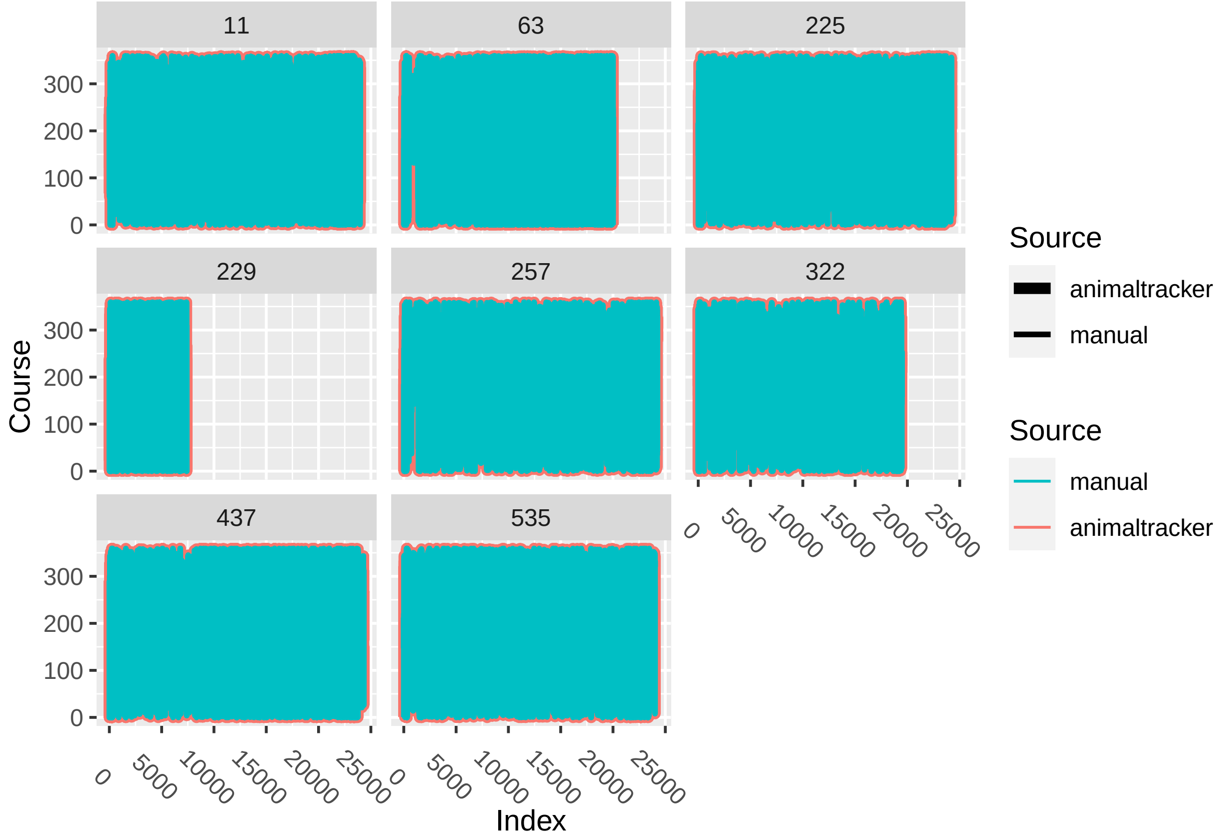
```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



Course by Cow

```
course_anitracker <- join %>%
  dplyr::select(Index, Cow, Course, CourseFlag.x) %>%
  dplyr::rename(Flag = CourseFlag.x) %>%
  dplyr::mutate(Source = "animaltracker")
```

```r
course_manual <- join %>%
  dplyr::select(Index, Cow, Course.y, CourseFlag.y) %>%
  dplyr::rename(Flag = CourseFlag.y,
                Course = Course.y) %>%
  dplyr::mutate(Source = "manual")

plot_data <- dplyr::bind_rows(course_anitracker, course_manual)

ggplot(plot_data, aes(x=Index, y=Course, group=Source, color=Source)) +
  geom_line(aes(size = Source)) +
  ylab("Course") +
  scale_color_discrete(guide = guide_legend(reverse = TRUE)) +
  scale_size_manual(values=c(2, 1)) +
  facet_wrap(vars(Cow)) +
  theme(axis.text.x = element_text(angle = -45))
```


Course
Index
Source
animaltracker
manual
Source
manual
animaltracker